Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 15:51:58
REFERENCE NO: 268

Author Names & Affiliations

- Nick Nystrom - Pittsburgh Supercomputing Center
- Ralph Roskies - Pittsburgh Supercomputing Center
- Robert Stock - Pittsburgh Supercomputing Center
- Sergiu Sanielevici - Pittsburgh Supercomputing Center

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

ACI service provider, representing the authors' research domains spanning the physical and life sciences, computer science, and computer engineering and their collective experience providing service across diverse research domains

Title of Submission

Need for NSF ACI centers and diverse hardware platforms to advance research frontiers and expand accessibility

Abstract (maximum ~200 words).

Research advances increasingly depend on sophisticated computer simulation, modeling, and analytics. Many projects are currently limited by aggregate system capacity and legacy software. Peer-reviewed requests exceed capacity by a factor of 3, and extensive software development effort is needed to couple, accelerate, and scale applications and to integrate them with instruments and analytics. Disruptive hardware and software technologies offer exciting potential for computational science and data analytics, yet they will require expertise and effort to exploit.

The single factor with the greatest potential impact across research challenges is sustaining multiple, strong ACI centers that provide: 1) experts who understand computational science, computer science, and advanced systems, and 2) diverse, heterogeneous, advanced hardware resources. Four or five ACI Centers are needed to operate a corresponding number of heterogenous computing systems, each (looking forward to next several years) providing in the range of 50-100 Pflops, where operations includes cross-cutting software development, user support, and training.

The research community would benefit from a model where Centers have some stability, e.g., in which each Center is periodically reviewed and, pursuant to successful review, extended for 5 or 10 years. Greater stability would increase Centers' impact and enable timely technical upgrades and better-planned system replacements.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

Advances in the life, physical, and planetary sciences, engineering, economics, and technology increasingly depend on sophisticated computer simulation, modeling, and analytics. Simulation and data-driven science are vital complements to experiment and observation, and researchers routinely apply a wide range of software. Yet rather than signaling saturation, we face an era of unprecedented opportunity. Increasing computational capability and capacity will support the coupling of applications, analytics, and data to study whole systems of interacting components. Individual components will be addressed at spatial and temporal scales vastly exceeding those possible today. In situ and standalone data analytics and visualization will let researchers understand and communicate results. Machine learning will continue to expand in importance, providing advanced analytics, coupling to HPC simulations to drive optimization, and contributing to novel force fields and other parameterizations.

With those opportunities come serious research challenges across multiple areas. In this response, we will not address specific, domain-specific research challenges, as those will be addressed by other investigators (including PIs at PSC). Rather, we focus on cross-cutting challenges that we observe on a national level and that require increased attention to enable U.S. research to advance. Those are as follows: coupling applications and data into complex workflows, extending applications to emerging computer architectures, scaling applications, and improving coordination of instruments and analytics. The single factor with the greatest potential impact across all of these research challenges is sustaining multiple, strong ACI centers that provide experts who understand computational science, computer science, and advanced systems; and diverse, heterogeneous, advanced hardware resources.

1.1. Coupling applications and data into complex workflows

Already the assembly of applications and large datasets into workflows is of very high importance. Examples include bioinformatic pipelines for genome and transcriptome assembly, coupled climate models such as the Community Earth System Model (CESM), and probabilistic seismic hazard analysis. Such workflows must manage tens to millions of program executions across multiple node types and multiple computational systems, manage provenance, and provide tools for collaboration and reproducibility. Ease of use is essential, because many users are not programmers or even comfortable with the Linux command line.

A special case of emerging workflows is the coupling of simulation and machine learning (ML), which offers considerable potential for areas such as accelerating optimizations, identifying rare events, and discovering more effective parameterizations. Going beyond today's workflows as mentioned above, coupling simulation and ML will require integration of very different application styles (e.g., MPI+X on an HPC subsystem and a deep learning framework on a GPU subsystem), possibly dynamic data assimilation and steering as processes run at different rates, and knowledge of the application domain, ML, and workflow frameworks.

Currently, there are many different workflow management frameworks; for example, Galaxy, Swift, Taverna, Kepler, Pegasus, and Apache Airavata. This situation poses multiple research challenges. First, there would be great benefit to identifying a minimal set of workflow management frameworks that can adequately address most application requirements. Second, those frameworks should be hardened and deployed uniformly across relevant advanced cyberinfrastructure, including sites such as today's XSEDE SPs and beyond, for example, to include data-intensive instruments (see #4, below). Third, important workflows must be expressed (some already are), hardened, and (importantly) optimized. For example, many bioinformatic pipelines have tremendous potential for optimization, which is critical because of the flood of data being generated by high-throughput sequencers. Finally, effective training must be provided to allow additional research communities to benefit and create a sustainable workforce.

1.2. Extending applications to emerging computer architectures

Over the coming decade, multiple disruptive technologies will drive profound diversification of computer architecture. We have already seen strong impact of GPU and manycore architectures, driven by the breakdown of Dennard scaling. Multiple programming models have emerged, such as OpenMP 4.0, CUDA, OpenACC, OpenCL, and ROCm. Currently there is a broad base of applications and libraries for GPUs and a growing base for manycore processors. However, those typically rely on different, independently-optimized code bases. Mainstream addition of forthcoming architectures such as Nervana and FPGAs will further increase heterogeneity. Another, even greater disruption will be due to the proliferation of new layers of memory, specifically, on-package HBM; NVRAM in DIMM form factors for extended RAM, caching, storage-class memory, and NVMe devices; and 3D NAND for NVMe and more traditional devices.

Recently, there have emerged cross-platform abstraction layers such as Kokkos and Raja. There are also plans for the evolution of programming languages, such as C++ executors. These abstraction layers, which leverage significant DOE investments toward exascale

computing, express data locality and concurrency at a higher level, offering significant potential for performance portability.

Two efforts are needed. First, applications in the NSF portfolio that will continue to be the greatest consumers of compute cycles should be evaluated for expression in a suitable abstraction framework, and, where appropriate, adapted. This would allow them to run efficiently on subsequent generations of GPUs, Xeon Phi, and other processors. Second, key applications' use of I/O and memory must be re-examined and optimized to exploit the new types of memory that will offer latencies and bandwidths closer to DRAM and with higher capacities. Again, software abstractions are emerging that will allow this to be done in a reasonable, forward-looking way.

1.3. Scaling applications

The research community will need ongoing effort to scale applications to compute platforms ranging from hundreds of petaflops to exaflops. We require scalability in both spatial and temporal scales. Increasing spatial resolution can reveal critical, qualitative changes in phenomenology which do not turn up at coarser scales (e.g., to understand the interaction of cloud microphysics on the development of severe storms), and similarly, finer temporal resolution can be necessary to reveal important details. We also require larger spatial models (e.g., to consider the coupling of different ground motions during an earthquake at the ends of a long bridge) and longer time scales (e.g., to understand multiscale defect initiation and failure in materials).

1.4. Improving coordination of instruments and analytics

Many types of instruments are emitting data at rates that are increasing faster than Moore's law. Current examples include high-throughput genome sequencers (130TB/year per Illumina HiSeq X sequencer), scanning confocal microscopy (~1TB/hour per instrument), and cryo-electron microscopy (~30TB/day per instrument). Some instruments are inherently community instruments, such as detectors on particle accelerators, while others generate data that is to be shared within a collaboration and/or (eventually) made public. Providing broad access to such data is democratizing, allowing researchers at other institutions to learn from, and contribute to, the knowledge base.

Maximizing the impact for research of data-intensive instruments requires distributed workflows (see 1.1), high-performance networking, sufficient computational capability for near-real-time analysis, and large-scale data management for curated community data collections.

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

ACI must be broadly usable solutions and let domain researchers focus on theory, algorithms, and collaboration. This response addresses the following areas: advanced computing systems, data sustainability, high-performance networking, ACI centers.

2.1. Advanced computing systems

Research applications require sufficient diversity of advanced computing platforms for optimal expression of algorithms, increasing capability and capacity and continuing and possibly expanding heterogeneity. The systems must have large, performant data management systems with external access (see 2.2) and high-performance networking to other computing systems, instruments, campuses, and labs (see 2.3).

Within this topic, the two greatest enablers of research are high capacity of reasonably powerful systems and hardware heterogeneity to support diverse applications.

Today, "reasonably powerful" means tens of petaflops, increasingly over time as processors of various types deliver ever-higher performance. Over the next few years, for the NSF community the sweet spot is likely to be around 50-100 Pflops rather than Exaflops. Even today, only a few "hero applications" can use the full leadership-class systems, requiring a broadening of the application base to include many workloads that allow only moderate scaling. Focusing system design to a level just below the maximum possible is more economical, yielding greater computational capacity for research as a whole. It is also more agile, allowing addition of new architectures at a faster pace, rather than overcommitting to technology that will be frozen in time. Increased capacity would be an immediate win for

proposed projects, which currently oversubscribe XSEDE resources by a factor of 3 based on merit review by the XRAC peer-review panel, with 200-300 proposals being submitted each quarter. Worthy of note is that PIs request allocations that they perceive as reasonable, not necessarily what their research actually requires. When asked to frame the problems they'd actually like to tackle, their capacity requirements are often substantially higher.

Substantial heterogeneity is needed to explore the rapidly-expanding options for hardware, memory, and storage technologies. For example, we already see the value of heterogeneity in systems such as Bridges at PSC, where commodity nodes provide substantial MPI+X capacity as well as Big Data frameworks, large-memory nodes uniquely serve genome sequence assembly and graph analytics, and GPU nodes drive many projects involving deep learning and simulation. Bridges complements the unique Anton 2 system at PSC, supplying necessary general-purpose computing for users to pre-equilibrate their MD systems. New, disruptive technologies and the ways that they will be incorporated into compute and storage systems will expand design possibilities. The research community needs access to a diversity of hardware architectures, supported by appropriate software and abstraction layers, to advance compute- and data-intensive science.

Often, significant economies could be generated by providing for technology refreshes of installed hardware. Considering the technology space of processor, accelerator, memory, and interconnect and the most promising ways to integrate them into systems, the frequencies at which new technologies emerge, and capacity requirements, the optimal number of reasonably powerful systems is 4 or 5. Options for technology refreshes would provide economy through continuation of some infrastructure, additional value to research, and increased productivity for users.

In addition, to provide early access to emerging technologies that may involve greater risk, smaller test-bed systems should be made available, with provision to expand those that prove valuable for the research community.

Capability-class systems, which we define as approximately an order of magnitude more powerful than the reasonably powerful systems, allow certain breakthroughs. However, their much higher cost must be balanced against rapid obsolescence, limited applications, and the high cost of scaling applications to that extreme. The perceived need for a capability-class system must also be balanced against the availability of DOE OSC systems to which NSF researchers can apply for access.

Budget permitting, one capability-class NSF system would be appropriate for extreme-scale research.

* ACI Systems vs. Cloud

A frequently-asked question is how advanced computing systems differ from commercial clouds. There are two key differentiators: ACI systems are purpose built to enable cutting edge research, while commercial clouds are built to develop business for commodity workloads; and ACI Centers provide extensive user support, while for commercial clouds there is only basic tech support.

Briefly, ACI systems as realized in ACI Centers offer the following pros and cons.
- Compute performance is high and predictable on HPC systems, versus variable and often low on virtualized commodity hardware.
- NSF ACI systems offer great value through their extensive software collections. This is an immediate productivity boost for users relative to spinning up their own instances and installing their own software stacks, especially where complicated dependencies make the task nontrivial.
- NSF ACI Centers offer great value through their expert user support; clouds offer only basic tech support for IaaS or PaaS. Users require help scaling, integrating, and running research applications, new users require introductory training, and many existing users require ongoing training in new methods.
- NSF ACI resources are available at no charge for open research. This is vital for onboarding new users, who cannot reliably estimate and hence budget their computing requirements, particularly data transfer. It is equally important for nontraditional fields, where there is often less funding, and for supporting university courses, graduate students, and workforce development.
- Commercial clouds offer an advantage where the priority is elasticity for burst processing of many small jobs. Many NSF-funded projects that require this already interoperate with commercial clouds.
- ACI systems and clouds both offer strong security and high reliability.

2.2. Data sustainability

The principal finding from the 2016 workshop at PSC on Best Practices for Data Infrastructure was that the research community needs a framework for data sustainability. This includes data management platforms, curation, mechanisms for data discovery and integration, and

connections to applications for working with the data. These factors are necessary to truly realize data management plans and to support data reuse and collaboration. Still needed is support to develop ACI that can be function effectively cross-program, robust, easily accessible to users, and supported by good documentation and training materials.

More fundamentally, the community needs a coherent policy from NSF regarding long term storage of data. Issues are who decides on what data is stored and for how long. The economics of storage is quite different from the economics of compute cycles. An unused compute cycle is lost; an allocated storage bit is lost to the rest of the community until it can be repurposed. Both are limited. Just as the community reviews applications for computer time, there needs to be a way to review proposals for storing data, and for how long. This could be revisited every year for each dataset. To date, there is no simple, accepted mechanism for a PI to request funding for storage of the data for the duration of the grant, and more importantly, no way to assure continued storage after the expiration of the grant.

2.4. ACI Centers

ACI Centers are necessary infrastructure in themselves to provide effective integration of systems and computational science experts. ACI Centers' additional roles are discussed in our response to Q3. The right number of centers scales with the number of systems, i.e., 4-6.

Question 3 Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

The greatest opportunity for sustaining and improving NSF ACI's benefit for research is organizational, specifically, going back to a model that recognizes Centers' role in providing both systems and deep human expertise.

Advanced computational resources are complex and exploit leading-edge, often first-of-their-kind technologies. Even when they contain some commodity components, they are integrated in innovative ways to deliver unique research capability. For example, PSC's Bridges was the first deployment in the world of the new Intel Omni-Path Architecture fabric, and PSC developed a custom topology to provide maximum value for new communities and data-intensive workloads. Bridges also pioneered use of OpenStack Ironic for bare-metal provisioning at scale, and Bridges' support of persistent databases, distributed services, and Big Data frameworks such as Spark and Hadoop extended the frontier of HPC. It was possible to conceptualize Bridges as a realistic system only because we could rely on PSC's very experienced systems staff to implement it. Experiences at other Centers are probably similar.

The current model, where Centers' existence is directly linked to competitions for new systems, poses grave risk for sustainability. Because of their unique skills, Center staff are aggressively recruited by Internet companies and vendors, who offer much higher salaries. The same applies to strong candidates who are entering the workforce. Retaining highly qualified staff at Centers, and attracting new talent to deal with attrition and new opportunities, requires a better model for sustainability. For example, the initial NSF Centers program was a 5-year program with a 5-year renewal. That model was key to initiating 3 of the Centers that exist today, and if re-enacted, it would allow refocusing of effort toward research results rather than survival.

Center staff are essential to the research challenges addressed in response to Question 1. For the software development challenges necessary to couple applications into workflows to simulate entire systems, Center staff, who bridge advanced resources, computer science, application domains, and training, are vital to successful implementation. We see this today through the large number of requests for XSEDE Extended Collaborative Support Service (ECSS), and the resulting co-authorship or acknowledgement of ECSS staff in users' publications. Yet, what can be accomplished through ECSS is quite limited and specific to individual projects, leaving unaddressed common infrastructure that cross-cuts domains and cases where more than 0.2 FTE over 1 year is required. Center staff can play a key role in more extensive, longer-term development of key software infrastructure.

Center staff provide analogous expertise for adapting and refactoring software to exploit emerging technologies, including the incorporation of recently-developed abstraction layers. Center staff are nearly unique in understanding both applications and advanced hardware, with their understanding of the latter often drawing on information obtained well prior to product availability. For example, we have seen systems where new accelerators have not been well-used because users were not able or interested in porting their applications. It is increasingly rare for applications groups to have internal expertise that spans their research domain, computer science, and HPC (CS programs rarely address HPC), and optimizing and scaling software does not, in itself, generate publications. Had Centers been more involved, for example, through "core" funding, greater value could have been had from that system investment, and applications could be better-

positioned now for the next generations of technology. Their efforts would leverage the ongoing investments of other agencies, which Center staff are already following, resulting in high return on investment.

Centers play an essential role in workforce development, training, STEM education, and to support industrial competitiveness. For example, looking only at Bridges, PSC has provided advanced data in HPC programming and Big Data to 3,697 people at 56 institutions, where 16 of those institutions are MSIs or in EPSCoR states. Participants from industry have attended at some of those sites. Over the past 5 years, PSC's MARC (Minority Access to Research Careers) program provided training and workforce development in bioinformatics to approximately 150 individuals at 25 institutions. PSC's BEST (Bioinformatics Education for Students) curriculum provides curriculum materials to high school teachers, and and GCode programs provide STEM+C training to high school students and teachers. To foster industrial competitiveness, PSC provides recommendations and expertise in topics such as HPC simulation and Big Data.

Centers need to provide a wide range of capabilities to cover the wide range of computing needs of the various application areas. This is best met by several Centers, supporting the 4-5 reasonably powerful, heterogenous systems, rather than one or two very large centers with very large systems, which would necessarily be more homogeneous and suffer rapid obsolescence. With requests for access to XSEDE computing resources exceeding available supply by a factor of three – after merit review – the need for ongoing access to a stable set of computing resources by researchers is manifest.

The current funding system for NSF centers requires all of the Centers to compete individually for the next single machine solicitation. If a Center fails to win the next machine when one of their current machines becomes outdated, the Center may have to close. This means that researchers cannot plan for the availability of resources, including user support. The effect on Center staff is chilling, resulting in the exit of valuable personnel and low morale.

The research community would benefit from a model where Centers have some stability. For example, each funded Center could reviewed periodically and, pursuant to successful review, extended, say for 5 or 10 years. Greater stability would increase Centers' impact by reducing the amount of time that is currently spent re-competing, enabling them instead to focus more clearly on enabling research. Greater stability would also enable more responsive technical upgrades, for example, budgeting for expanding capacity using new but similar technology part way into a system's lifetime, and better-planned system replacements.

Operations funding for operations and management, including system administration, system-specific user support, outreach, training, and power and cooling, also needs to be increased. The current model of 20% per annum of the acquisition cost does not realistically reflect those costs. Increasing the O&M budget would allow significantly higher value to be obtained from the investment in the acquisition.

Consent Statement

- "I hereby agree to give the National Science Foundation (NSF) the right to use this information for the purposes stated above and to display it on a publically available website, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode)."